Quick Deploys

Quick Deploys lets you deploy custom Endpoints of popular models with minimal configuration.

You can find Quick Deploys and their descriptions in the Web interface.

How to do I get started with Quick Deploys?

You can get started by following the steps below:

- 1. Go to the Serverless section in the Web interface.
- 2. Select your model.
- 3. Provide a name for your Endpoint.
- 4. Select your GPU instance.
 - i. (optional) You can further customize your deployment.
- 5. Select **Deploy**.

Your Endpoint Id is now created and you can use it in your application.

Customizing your Functions

To customize AI Endpoints, visit the RunPod GitHub repositories. Here, you can fork the programming and compute model templates.

Begin with the worker-template and modify it as needed. These RunPod workers incorporate CI/CD features to streamline your project setup.

For detailed guidance on customizing your interaction Endpoints, refer to Handler Functions.

Edit this page